Directional Hallucinations: Ideological Drift in News-Grounded LLM Question Answering

Chendi Wang , Liam Cunningham , Tom Yishay , Jieying Chen
Vrije Universiteit Amsterdam
chendi.wang@vu.nl, l.e.cunningham@student.vu.nl, t.yishay3@student.vu.nl, j.y.chen@vu.nl

Abstract

Large language models (LLMs) are increasingly used to answer questions about political information, including in election-adjacent information settings where factual errors and ideological distortions are high-stakes. We present a reproducible measurement framework that treats hallucinations, unsupported statements in document-grounded QA, as diagnostic signals of ideological drift. Using 21,727 expert-labeled U.S. political news articles from QBias spanning left, center, and right sources, we (i) generate an article-specific question, (ii) elicit document-grounded answers from three open-weight LLMs and one proprietary model, (iii) detect sentence-level hallucinations via reference-based comparison, (iv) classify the ideological valence of hallucinated sentences with a fine-tuned stance classifier, and (v) probe output logits to relate token-level uncertainty to hallucination and drift. Hallucination rates vary substantially across models and concentrate in contentious topics, while source-ideology differences in hallucination frequency are modest. In contrast, hallucination content exhibits robust leftward drift: a majority of hallucinated sentences are classified as left-leaning, including among hallucinations generated from right-leaning sources. Logit-level analysis shows hallucinations arise in high-entropy generation contexts, and in some models uncertainty also predicts leftward drift, consistent with an "uncertainty $\rightarrow$ guessing" mechanism. We discuss implications for auditing AI-mediated political information and for designing safeguards in election-relevant deployments.

1 Societal Problem and Real-World Impact

Societal problem. As LLMs are integrated into search, chat assistants, and news interfaces, they increasingly mediate political information access. In civic and electoral contexts, hallucinated or ideologically skewed answers can misinform voters, distort issue salience, and erode trust in institutions and platforms [Weidinger *et al.*, 2021; Bender *et al.*, 2021; Allcott and Gentzkow, 2017; Vosoughi *et al.*, 2018]. Recent evidence shows conversational AI can be politically persuasive in controlled settings [Hackenburg *et al.*, 2025; Lin *et al.*, 2025; Hackenburg and Margetts, 2024], raising concerns about downstream effects when AI systems generate incorrect or ideologically slanted claims.

Real-world impact. Our output is an auditable, model-agnostic evaluation workflow that can be used by platforms and public institutions to stress-test LLM deployments for civic information. The pipeline produces: (i) hallucination prevalence by source ideology and topic, (ii) directional drift metrics with uncertainty intervals, and (iii) uncertainty-linked risk indicators usable for abstention or warning policies (e.g., "do not answer when uncertainty exceeds threshold") [Kalai *et al.*, 2026].

2 Introduction

LLMs are increasingly deployed as interfaces to political and civic information, often in *grounded* settings where a system is asked to answer a question using a provided document (e.g., a news article) or retrieved sources. Even in such settings, models can produce hallucinations, i.e. unsupported statements that appear confident and coherent [Huang *et al.*, 2025; Ji *et al.*, 2024; Kalai *et al.*, 2026]. Separately, a growing literature documents political bias and value misalignment in LLM outputs, frequently showing systematic skews that vary by topic and evaluation design [Rettenberger *et al.*, 2025; Bang *et al.*, 2024; Chen *et al.*, 2026; Navigli *et al.*, 2023; Motoki *et al.*, 2025; Gallegos *et al.*, 2024].

Core question. What happens when these two failure modes intersect? When an LLM loses grounding and begins to "fill in" missing information, is the resulting hallucinated content ideologically *directional*? If so, that would imply hallucinations are not merely reliability failures, but also *bias-revealing events* that surface latent priors.

Why this matters for social good. In civic contexts (e.g., election information access), an LLM that hallucinates could introduce ideologically slanted misinformation even when the underlying source is ideologically opposed or neutral, undermining fairness, trust, and informed decision-making [Weidinger *et al.*, 2021; Bender *et al.*, 2021; Allcott and Gentzkow, 2017].

Contributions. Our contributions are fourfold:

- **Measurement framework:** a transparent, reproducible pipeline for auditing *ideological directionality* of hallucinated content in news-grounded QA.
- **Empirical finding:** hallucination rates vary by model and are modestly higher for left sources, but hallucination content shows robust leftward drift even from right sources.
- **Topical risk profiling:** drift and hallucination vulnerability are amplified on contested topics (including election-relevant issues).
- **Mechanistic evidence:** logit-level uncertainty signals predict hallucination events; uncertainty analyses help distinguish "guessing" from source-conditioned generation, and clarify whether drift reflects uncertainty or priors.

3 Related Work

Political bias in language models. Multiple approaches quantify political bias in LLMs, including ideology tests, policy questionnaires, and comparisons to human polling distributions [Bang *et al.*, 2024; Santurkar *et al.*, 2023; Motoki *et al.*, 2025; Rozado, 2025]. Bias is not monolithic: it varies by topic and evaluation protocol [Motoki *et al.*, 2025; Navigli *et al.*, 2023; Gallegos *et al.*, 2024]. These findings motivate bias auditing in *realistic tasks* rather than only prompt-based questionnaires.

Hallucinations and grounded generation. Hallucinations in QA and summarization are widely documented [Huang *et al.*, 2025; Ji *et al.*, 2024; Zhang *et al.*, 2023]. Recent work argues hallucinations arise partly because training and evaluation reward guessing over abstention [Kalai *et al.*, 2026]. Detection methods include human evaluation, entailment-style scoring, and learned annotators [Gu *et al.*, 2024]. Our work uses a reference-based hallucination detector to identify unsupported sentences in grounded QA outputs.

Uncertainty and abstention. Confidence estimates in neural models are often miscalibrated [Guo *et al.*, 2017]. Selective prediction and abstention policies can mitigate harm when uncertainty is high [Geifman and El-Yaniv, 2017]. In LLM contexts, logit-derived metrics (entropy, margins) provide interpretable uncertainty signals that can be connected to hallucination risk [Kalai *et al.*, 2026; Jiang *et al.*, 2021].

Democracy, elections, and conversational AI. AI systems can influence political attitudes and persuasion dynamics [Hackenburg *et al.*, 2025; Lin *et al.*, 2025; Bai *et al.*, 2025]. These results heighten the importance of auditing not only overt persuasion, but also *misinformation pathways* produced through hallucinated civic and political information. This question carries significant societal implications. If hallucinations systematically inject political bias into news-grounded QA, users assuming AI neutrality may unknowingly consume ideologically skewed misinformation. Such content could reinforce echo chambers [Cinelli *et al.*, 2025], distort public perception, and erode trust in both AI systems and the information they convey, particularly concerning given LLMs' growing role in search and content moderation [Noels *et al.*, 2025]. The problem is compounded by the fact that hallucinated content often appears fluent and confident, making it difficult for users to distinguish fabricated claims from accurately summarized information.

4 Data and Task

Dataset. We use the QBias corpus [Haak and Schaer, 2023], which contains 21,747 U.S. political news articles collected from AllSides balanced news headline roundups (June 2012–November 2022) and labeled by expert annotators as *left*, *center*, or *right*. After cleaning (duplicate removal, missing text), our final set contains $N = 21727$ articles with counts: left=10261, right=7214, center=4252. We retain topical tags provided by QBias (e.g., Elections, Immigration, Gun Control), enabling topic-level risk analysis. This imbalance reflects the underlying distribution of partisan sources and is accounted for through normalization in our analysis. The temporal scope (2012–2022) captures a particularly relevant period for studying political discourse, encompassing three presidential administrations, multiple election cycles, and the emergence of "fake news" as a prominent concern.

Task: news-grounded question answering. We evaluate *grounded QA* rather than generic summarization to reduce degrees of freedom: each model must answer a question using only the provided article as evidence. For each article:

1. **Question generation:** generate a single neutral question capturing the core claim of the article.
2. **Grounded answering:** provide a short answer in sentence form using only the article content.

This yields a controlled setting where sentence-level hallucination detection is well-defined.

5 Methodology

Figure 1 illustrates the end-to-end pipeline of our methodology, from question generation to hallucination analysis and ideological assessment.

5.1 Models and Generation

We evaluate three instruction-tuned open-source LLMs and one proprietary LLM: *Llama 3 8B Instruct*, *Mistral 7B v0.3 Instruct*, *Deepseek 7B Chat* and *GPT-4o-Mini*. For each article, we generate a single neutral question using Llama 3 and then have each of the four models answer the same question using the full article as context (Appendix: prompts and decoding parameters). Using a single shared question per article ensures comparability across models and mitigates question-content confounds. We segment each model answer into sentences; all hallucination and ideology analyses are conducted at the sentence level.

5.2 Hallucination Detection (Sentence-Level)

We apply ANAH-v2 [Gu *et al.*, 2024] to compare each answer sentence against the corresponding source article and assigning one of three labels: *ok* (supported), *unverifiable* (not supported), or *contradictory* (conflicts with the article). In all subsequent analyses, we collapse *unverifiable* and *contradictory* into a single hallucination indicator. We define a binary hallucination indicator $\mathbb{H} \in \{0, 1\}$ at the sentence level as

$$\mathbb{H} = \mathbb{I}(\ell \in \{\textit{unverifiable}, \textit{contradictory}\}),$$

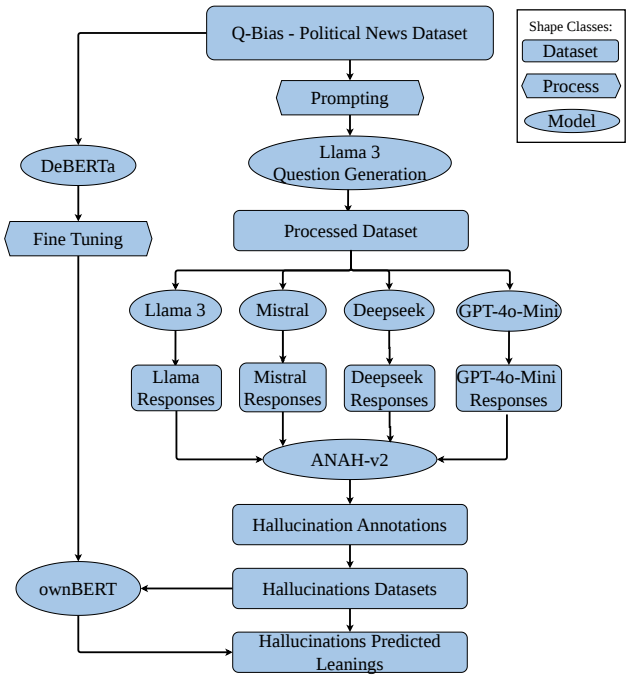

Figure 1: Pipeline overview: (1) question generation, (2) news-grounded answer generation, (3) sentence-level hallucination detection, (4) ideology classification of hallucinated sentences (binary left/right), and (5) logit-based uncertainty probing.

where $\mathbb{I}(\cdot)$ is the indicator function and ℓ denotes the ANAH-v2 sentence label.

5.3 Ideology Classification of Hallucination

To label the ideology of hallucinated content, we fine-tune DeBERTa-v3 [He *et al.*, 2021] on QBias article text for **binary** left/right classification. We adopt the binary setup because ternary (left/center/right) classification is substantially less reliable on QBias (F1 = 0.62 vs. 0.74), which would introduce additional noise and reduce the robustness of drift estimation. The exclusion of the "center" articles is therefore a methodological choice driven by classifier reliability, rather than an assumption about the underlying ideological structure. We downsample the larger class so that the left and right training sets are balanced at $N_{\min} = \min(N_{\text{left}}, N_{\text{right}})$ articles per class. The resulting classifier achieves test macro-F1 = 0.74.

We apply this classifier to hallucinated sentences and interpret its outputs as *hallucination-content ideology*. Because the classifier is binary, all **directional drift analyses (RQ2)** are restricted to: source articles labeled left or right (center excluded), and hallucinated sentences classified as left or right (binary output).

We nevertheless include **center** sources in **RQ1** hallucination-rate and topic-risk profiling.

5.4 Directional Drift Metrics and Hypothesis Tests

Let $s \in \{L, R\}$ denote the ideology of source-article ideology (restricted to left/right for drift analyses) and $\hat{y} \in \{L, R\}$ denote the predicted ideology of a hallucinated sentence.

We report the following metrics:

- *Alignment rate.* $\Pr(\hat{y} = s \mid \text{hallucination}, s \in \{L, R\})$.
- *Leftward drift from right sources.* $\Pr(\hat{y} = L \mid \text{hallucination}, s = R)$.
- *Overall left-share among hallucinations.* $\Pr(\hat{y} = L \mid \text{hallucination}, s \in \{L, R\})$.

To test directionality, we conduct one-sided binomial tests of $\Pr(\hat{y} = L)$ against a symmetric null $p_0 = 0.5$ and report exact 95% confidence intervals and Cohen's h effect sizes.

5.5 Logit-Level Uncertainty Probing (RQ3)

To investigate underlying generation mechanisms, we extract final-layer token logits during decoding and compute per-sentence uncertainty measures. For each token with predictive distribution $p(\cdot)$, entropy is defined as:

$$H(p) = -\sum_{v} p(v) \log p(v).$$

We aggregate uncertainty at the sentence level by computing the mean token entropy and include token count as a control for sentence length. We restrict the logit-level analysis to the three open-weight models: for these we compute entropy over the full output distribution, whereas GPT-4o-Mini exposes only its top-5 token log-probabilities through the API, yielding entropy that is not comparable in scale; we report a separate probe of its logits in Section 6.3.

T1 (uncertainty → hallucination). We fit logistic regression models predicting sentence-level hallucination from mean token entropy, while controlling for sentence length (token count). We report odds ratios, 95% confidence intervals(CIs), and ROC–AUC.

T2 (uncertainty → drift, within hallucinations). Restricting to hallucinated sentences, we fit logistic regressions predicting whether hallucinated content is classified as left-leaning (vs right-leaning) from mean entropy and token count. We estimate models separately by LLM to characterize heterogeneity in the uncertainty–drift relationship. Table 1 summarizes our experimental setup in the paper.

Component	Details
Dataset	QBias (cleaned): 21,727 articles
Time period	June 2012 – November 2022
LLMs evaluated	Llama 3 8B, Mistral 7B, Deepseek 7B, GPT-4o-Mini
Hallucination detector	ANAH-v2 (∼90% accuracy)
Stance classifier	ownBERT (DeBERTa-v3, F1 = 0.74)

Table 1: Experimental setup summary.

6 Results

6.1 RQ1: Do Hallucination Rates Vary by Source Ideology and Topic?

Overall hallucination rate. To quantify model behaviour under grounding failure, we first examine the overall volume and type of hallucinations each system produces. Understanding baseline hallucination rates is essential because

it determines how much ungrounded content contributes to downstream ideological drift.

Substantial differences emerge across models despite similar parameter counts. DeepSeek produced the highest number of hallucinations (7,090 sentences), nearly doubling the counts of Llama (3,702) and Mistral (4,475), while GPT-4o-Mini generated the fewest (1,923). This variation suggests that hallucination propensity depends not only on parameter count but also heavily on model architecture and training procedures.

The substantial variation in hallucination rates across models with similar parameter counts warrants further examination. In percentage terms, hallucination rates also differ significantly across models ($\chi^2 = 2082.20$, $p < 0.001$). Deepseek exhibits the highest hallucination rate (21.3%), substantially exceeding Llama (14.1%), Mistral (14.2%) and GPT-4o-Mini (7.8%). Pairwise comparisons reveal that Deepseek differs significantly from both Llama and Mistral ($p < 0.001$), while Llama and Mistral show identical rates ($p = 0.72$). GPT-4o-Mini also differs significantly from all three open-weight models (all $p < 0.001$), exhibiting the lowest hallucination rate. Llama 3's relatively low hallucination rate within the open-weight camp may partially reflect its use for question generation—questions it formulated may align better with its internal knowledge organization, providing a slight advantage. However, even accounting for this potential confound, Deepseek's hallucination rate represents a substantial increase over Llama, suggesting meaningful architectural or training differences beyond our experimental design. Moreover, GPT-4o-Mini's superior performance despite not benefiting from the same question-generation advantage suggests that its lower hallucination rate reflects genuine differences in model capabilities rather than an artifact of the evaluation setup. This pattern suggests that Deepseek's elevated hallucination propensity reflects a model-specific property, potentially related to training data composition or architecture, rather than task difficulty, since Llama and Mistral perform equivalently on the same inputs. Notably, all three open-weight models were instruction-tuned, yet their hallucination profiles differ markedly. This suggests that instruction tuning, while improving task compliance, does not uniformly address factual grounding.

By source ideology. A natural next question is whether models' hallucination frequency varies by one ideological side. After normalizing for dataset imbalance, the three open-weight models exhibit hallucinations at substantially higher rates for left-leaning source articles: Deepseek (22.5% left vs. 19.7% center vs. 20.7% right), Llama (15.0% vs. 12.8% vs. 13.4%), and Mistral (15.1% vs. 12.4% vs. 14.0%). This statistically significant pattern ($p < 0.001$) is remarkably consistent across the three models, suggesting that left-leaning content may be systematically more challenging for them to ground accurately. In contrast, GPT-4o-Mini shows virtually no variation across source ideology (7.9% vs. 7.7% vs. 7.8%; $p = 0.927$). This suggests that the ideology-linked variation in hallucination rates is concentrated among the open-weight models rather than being a universal property of the task. Potential explanations include differences in linguistic complexity, topic distribution, gaps in training data representation, or model-specific factual grounding behavior.

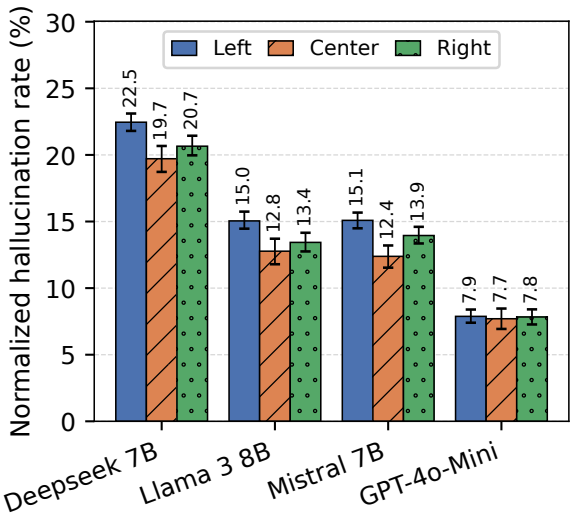

Figure 2: Normalized hallucination proportions by source leaning.

By topic. Beyond overall frequency, hallucination propensity may vary by topical domain. Political news spans diverse issue areas that differ in polarity, factual stability, and narrative complexity. To illuminate where models struggle most, we analyze hallucination rates by topic in Figure 3.

Figure 3 shows hallucination rates by topic, normalized to baseline topic distribution in the dataset. Hallucinations cluster around contested political topics. Across all models, the top hallucination-prone topics include presidential politics, presidential elections, healthcare, immigration and gun control. These topics share common characteristics: they feature rapidly evolving information, contested factual claims, and strong partisan framing in media coverage. This suggests that training data composition plays a role—topics with abundant but conflicting information may leave models with uncertain associations, increasing hallucination likelihood.

6.2 RQ2: Is Hallucinated Content Ideologically Directional?

Alignment vs. misalignment. After understanding where hallucinations occur, we next assess whether hallucinated content stays consistent with the ideological direction of the source article. This alignment analysis reveals how often hallucinations reinforce versus diverge from the original narrative. Because our stance classifier is binary for reliability, RQ2 drift analyses are restricted to source articles labeled left/right and hallucinated sentences classified as left/right.

To quantify ideological alignment, we compare own-BERT's predicted leaning for each hallucinated sentence with the expert-annotated leaning of its source article. A sentence is considered *aligned* if the labels match (left–left or right–right) and *misaligned* otherwise. Figure 4 reports the proportion of hallucinated sentences whose predicted political leaning matches the orientation of the source article. Across all models, only 56–58% of hallucinations preserve the ideologi-

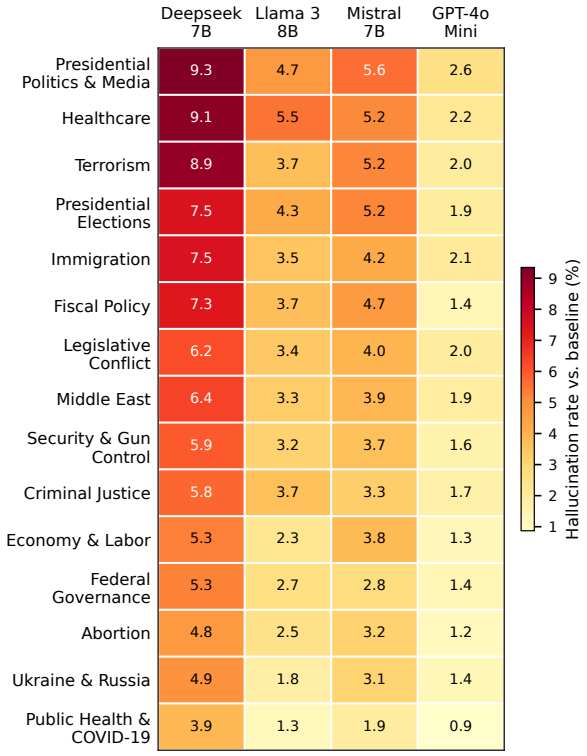

Figure 3: Hallucinated topics by model (normalized by baseline topic frequency).

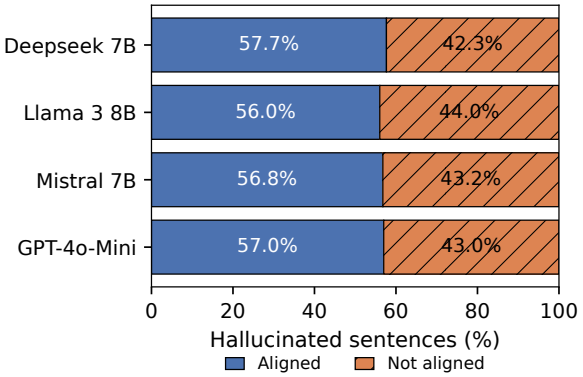

Figure 4: Ideological alignment of hallucinations: proportion matching source orientation (Aligned) vs. diverging (Not Aligned).

Condition	**% Left**	**95% CI**	p**-value**	h
Overall	67.5	[66.7, 68.3]	< 0.001	0.36
Deepseek 7B Chat	69.8	[68.6, 71.0]	< 0.001	0.41
Mistral 7B v0.3 Instruct	67.0	[65.4, 68.5]	< 0.001	0.35
Llama 3 8B Instruct	64.2	[62.5, 65.9]	< 0.001	0.29
GPT-4o-Mini	66.9	[64.5, 69.2]	< 0.001	0.34
Right sources only	63.2	[61.9, 64.5]	< 0.001	0.27

Table 2: Leftward drift in hallucinated outputs. Entries report the proportion of left-leaning predictions among hallucinated responses classified as either left or right. p-values are from one-sided binomial tests against a null of equal left/right probability ($p_0 = 0.5$). Models show drift with medium effect sizes (Cohen's $h > 0.25$).

cal direction of their source—barely above chance. Approximately 43% of fabricated content introduces perspectives *absent* from the original article. Differences across models are minimal (56.0% for LLaMA, 56.8% for Mistral, 57.7% for DeepSeek and 57.0% for GPT), indicating that while hallucination *frequency* varies substantially across models, ideological alignment remains largely invariant.

Directional drift. The alignment analysis reveals that roughly 43% of hallucinations diverge from their source orientation, but this alone does not indicate whether drift is random or directional. The critical question is: in which direction do misaligned hallucinations drift? If errors were symmetric, we would expect roughly equal proportions of left/right misclassifications. Systematic asymmetry, by contrast, would indicate an underlying ideological bias in ungrounded generation. To measure the directionality of misaligned hallucinations, we compare the predicted ideological leaning of each hallucination against its source article.

Figure 5 shows the confusion matrices comparing source article orientation (rows) with predicted hallucination orientation (columns). All four models exhibit systematic leftward drift: regardless of source ideology, hallucinated content is predominantly classified as left-leaning. Examining the overall proportion of left-leaning predictions among all predictions: Deepseek shows the strongest skew (69.8% of all hallucinations classified as left), followed by Mistral (67.0%), GPT-4o-Mini (66.9%) and Llama (64.2%).

Critically, this leftward skew persists even when doing news-grounded QA of right-leaning articles. For right-leaning sources, the proportion of hallucinations classified as left-leaning is: Deepseek 65.5%, Mistral 62.8%, GPT-4o-Mini (61.9%) and Llama 60.3%. This asymmetry confirms that hallucinations do not introduce random noise, as they systematically inject left-leaning content regardless of source orientation. The consistency of this pattern across four independently developed models suggests a common underlying cause, likely related to the composition of English-language training corpora.

Given the frequency asymmetry above, one might hypothesize that hallucination political bias simply reflects the higher volume of hallucinations from left-leaning sources. However, this explanation fails: even when models hallucinate on right-leaning source articles, the hallucinated content still drifts leftward. This demonstrates that leftward content bias is independent of source ideology, and the models inject left-leaning content regardless of what they are reading. Table 2 presents binomial test results.

If question generation confounded results by biasing content toward Llama's priors, Llama should exhibit the strongest leftward drift. Instead, Llama shows the *weakest* drift (64.2% vs. Deepseek's 69.8%; $\chi^2 = 30.51$, $p < 0.001$). Furthermore, Llama and Mistral show statistically identical hallucination rates (14.1% vs. 14.2%, $p = 0.72$), suggesting no unique Llama advantage. These patterns argue against the

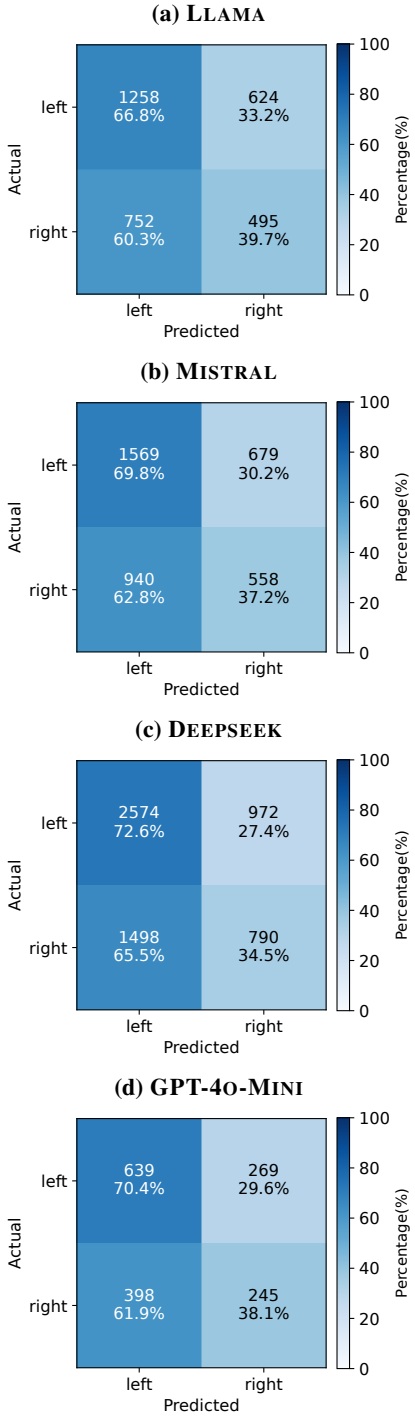

Figure 5: Confusion matrices: source article orientation (rows) vs. predicted hallucination orientation (columns).

QG confound driving our findings.

6.3 RQ3: Do Uncertainty Patterns Explain Hallucination and Drift?

T1: Uncertainty predicts hallucination. Logit-level analysis indicates that hallucinations arise in high-uncertainty generation contexts. Mean token entropy is substantially higher for hallucinated sentences ($M = 0.639$) than for correct sentences ($M = 0.323$), yielding a difference that is both statistically significant (Welch's $t = 100.3$, $p < 0.001$) and practically large (Cohen's $d = 1.03$).

To quantify the predictive relationship, we fit a logistic regression predicting hallucination from mean entropy while controlling for response length (number of tokens). Each unit increase in mean entropy multiplies hallucination odds by 40.4 (95% CI: [37.6, 43.3], $p < 0.001$). The model achieves ROC-AUC = 0.78, indicating good discriminative ability. This supports an "uncertainty $\rightarrow$ guessing" mechanism consistent with [Kalai *et al.*, 2026].

T2: Uncertainty-ideology relationship. We also examined whether higher uncertainty predicts leftward drift within hallucinations, using logistic regression to control for response length. Models differ markedly:

Deepseek shows a strong uncertainty-ideology relationship: each unit increase in mean entropy multiplies the odds of left-leaning hallucination content by 2.32 (95% CI: [2.02, 2.67], $p < 0.001$). Leftward bias emerges specifically in high-uncertainty contexts. Llama shows a moderate relationship in the same direction (OR = 1.47, 95% CI: [1.14, 1.89], $p = 0.003$). Mistral shows no significant uncertainty-ideology relationship (OR = 0.71, 95% CI: [0.50, 1.02], $p = 0.065$), indicating that Mistral's leftward bias operates independently of model confidence, i.e. the bias is present regardless of uncertainty level.

This heterogeneity suggests different bias mechanisms: Deepseek and Llama's biases activate under uncertainty, while Mistral's bias appears more uniformly embedded. This indicates that while all models share leftward-biased priors, the mechanism differs across models.

Mechanistic interpretation. Our results adjudicate between two hypotheses about how leftward bias manifests. Under H1 (uncertainty-gated priors), leftward drift becomes more likely as uncertainty increases. Under H2 (uncertainty-invariant priors), drift is present regardless of uncertainty level. Our results support H1 for Deepseek and Llama: within hallucinations, higher entropy significantly increases the odds that hallucinated content is classified as left-leaning. In contrast, Mistral shows no positive uncertainty–drift relationship, consistent with H2 or with mechanisms not captured by entropy alone, suggesting a more uniformly embedded prior that neither hypothesis fully explains. We therefore interpret uncertainty as a robust trigger for hallucination, and (for some models) a trigger for directional drift; identifying the source of leftward priors (e.g., training data composition or alignment) remains outside the scope of these analyses.

This heterogeneity implies that bias mitigation strategies may need to be model-specific: uncertainty detection may help for Deepseek and Llama, while Mistral may require interventions at the training or architectural level. These findings also explain existing observations that AI produces more errors when advocating right-leaning positions [Lin *et al.*, 2025] generating right-leaning content requires overcoming the model's latent leftward prior, increasing error susceptibility. While our results illuminate the *trigger mechanism*, they do not identify the *source* of leftward priors, which likely re-

flects training data composition.

Closed-model probe. We also applied the logit analysis to GPT-4o-Mini, but its API exposes only the top-5 token log-probabilities. The resulting entropy is sharply compressed with max sentence-level entropy 0.70 nats versus 3.3–3.4 for the open-weight models, and the separation between hallucinated and correct sentences nearly vanishes (0.04 nats vs. 0.33–0.38), leaving its uncertainty–drift estimate uninformative (OR = 1.58, 95% CI: [0.62, 4.03], $p = 0.343$). We therefore exclude GPT-4o-Mini from the RQ3 estimates above. This limitation is itself relevant to auditing: a deployed proprietary model's uncertainty mechanism cannot be probed this way without fuller access to its output distribution.

7 Discussion

Directional hallucinations as structured errors. Our results separate two phenomena that are often conflated:

- **Reliability (how often models hallucinate):** models hallucinate more frequently on left-leaning sources, but the difference is not large across left/center/right source material in our grounded QA setup.
- **Ideological structure of errors (what they hallucinate):** systematically left-skewed hallucination content, even when the source is right-leaning. This finding aligns with documented left-leaning tendencies in LLM outputs [Rozado, 2025; Lin *et al.*, 2024; Chen *et al.*, 2026].

This dissociation matters for social-good deployments: even if overall hallucination rates are reduced, the remaining errors may still be directionally biased.

Mechanisms: uncertainty, guessing incentives, and priors. Logit-level evidence supports that hallucinations concentrate in high-uncertainty contexts, aligning with the argument that current training/evaluation regimes reward "guessing" [Kalai *et al.*, 2026]. Yet leftward drift persists even when uncertainty patterns differ by model, consistent with the view that hallucinations expose asymmetric priors embedded during pretraining and alignment (e.g., corpus composition, RLHF preferences, or safety tuning) [Navigli *et al.*, 2023; Gallegos *et al.*, 2024; Santurkar *et al.*, 2023].

Election-specific implications. In election-adjacent contexts, both *factual error* and *perceived partisan skew* can reduce trust, even when systems are intended to be neutral. Our topic-level profiling provides a practical pathway: models can be audited and constrained more aggressively on election-relevant topics (e.g., impose abstention at high uncertainty, require citations, enforce retrieval grounding) [Geifman and El-Yaniv, 2017; Guo *et al.*, 2017; Lewis *et al.*, 2020].

8 Limitations

Classifier limitations. Our classifier (F1 = 0.74) may introduce noise, but random errors bias toward null, making our significant findings conservative. We cannot fully rule out differential accuracy across ideologies; future work should examine classifier calibration by ideology.

Data contamination. QBias articles (2012–2022) may appear in training corpora. However, memorization would *reduce* hallucination, not bias direction. Consistent leftward drift across models with different training data (including Deepseek) suggests contamination is not the primary driver.

Detection validity. ANAH-v2 was not validated specifically on political news. While ANAH-v2 was validated on data containing substantial political content (37% Politics/Military), our U.S. partisan news sources may differ in style from the validation corpus.

Question generation confound. Using one model for question generation may advantage that model in answering. If Llama-generated questions biased content leftward, Llama should show the strongest effect. Instead, Llama shows the *weakest* drift (64.2%), arguing against this confound.

Logit access for closed models. Our logit-level uncertainty analysis (RQ3) requires access to the full output distribution, which closed APIs do not expose; as quantified by the closed-model probe in Section 6.3, this precludes a comparable analysis for GPT-4o-Mini, so RQ3 is restricted to the open-weight models. GPT-4o-Mini's RQ1 and RQ2 results, which do not rely on logits, are unaffected.

9 Implications for AI-Mediated Information Systems

These findings have broad implications for the design, deployment, and governance of AI systems that mediate political information:

For users and society: As LLMs become integrated into news consumption workflows, users may unconsciously internalize politically biased misinformation. Even when source material is balanced or right-leaning, hallucinations can inject left-leaning perspectives, potentially reinforcing echo chambers and contributing to societal polarization—particularly concerning given assumptions of AI neutrality. Users who rely on AI summaries without consulting original sources may develop systematically skewed understandings of political issues.

For platform design: AI-powered QA systems should implement hallucination-aware content warnings and provide transparency about model tendencies. Factual grounding through retrieval-augmented generation [Niu *et al.*, 2024] may help, but our results suggest that when grounding fails, systematic biases emerge. Platforms might consider displaying confidence indicators or providing links to source material for verification.

For AI governance: Policymakers should consider establishing accuracy and neutrality standards for AI systems used in politically sensitive contexts. Our methodology provides a template for auditing political bias in AI-generated content that could be applied to new models as they are released. Recent work has also revealed linguistic bias in governmental documents [de Swart *et al.*, 2025], suggesting that bias can emerge in both source texts and AI-generated content.

For model development: Bias mitigation efforts should be topic-aware, as our analysis reveals that contested issues amplify partisan biases. Simply reducing overall hallucination rates may not address the systematic ideological skew in remaining hallucinations. Developers might consider balanced training data curation or targeted fine-tuning on politically sensitive topics.

Ethical Statement

This research relies exclusively on publicly available, annotated news records. The dataset contains no personally identifiable information and does not involve private individuals. No human participants were recruited or directly engaged in the study; consequently, informed consent was not required.

Acknowledgements

This work was supported by the Dutch Research Council (NWO) through the project *PoliBiasEU: A Scalable and Multilingual Benchmark for Political Bias Detection in Large Language Models* (project no. ZZKNS44563) and by the Dutch Ministry of Education, Culture and Science (OCW) in the scope of the project *Navigating the Storm: European Political Contestation in Geopolitical Transformation (NEST)*, and partially supported by the Network Institute, Vrije Universiteit Amsterdam, through the Academy Assistants Program.

References

[Allcott and Gentzkow, 2017] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236, 2017.

[Bai *et al.*, 2025] Hui Bai, Jan G Voelkel, Shane Muldowney, Johannes C Eichstaedt, and Robb Willer. Llm-generated messages can persuade humans on policy issues. *Nature Communications*, 16(1):6037, 2025.

[Bang *et al.*, 2024] Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. Measuring political bias in large language models: What is said and how it is said. *arXiv preprint arXiv:2403.18932*, 2024.

[Bender *et al.*, 2021] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[Chen *et al.*, 2026] Jieying Chen, Karen de Jong, Andreas Poole, Jan Burakowski, Elena Elderson Nosti, Joep Windt, and Chendi Wang. Uncovering political bias in large language models using parliamentary voting records. *arXiv preprint arXiv:2601.08785*, 2026.

[Cinelli *et al.*, 2025] Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. The echo chamber effect on social media. *ResearchGate*, February 2025.

[de Swart *et al.*, 2025] Milena de Swart, Floris den Hengst, and Jieying Chen. Detecting linguistic bias in government documents using large language models. In Guodong Long, Michale Blumestein, Yi Chang, Liane Lewin-Eytan, Zi Helen Huang, and Elad Yom-Tov, editors, *Proceedings of the ACM on Web Conference 2025, WWW 2025, Sydney, NSW, Australia, 28 April 2025- 2 May 2025*, pages 5034–5044. ACM, 2025.

[Gallegos *et al.*, 2024] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.

[Geifman and El-Yaniv, 2017] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30, 2017.

[Gu *et al.*, 2024] Yuzhe Gu, Ziwei Ji, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. Anah-v2: Scaling analytical hallucination annotation of large language models. *Advances in Neural Information Processing Systems*, 37:60012–60039, 2024.

[Guo *et al.*, 2017] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[Haak and Schaer, 2023] Fabian Haak and Philipp Schaer. Qbias - A Dataset on Media Bias in Search Queries and Query Suggestions. In *Proceedings of the 15th ACM Web Science Conference 2023*, pages 239–244, April 2023.

[Hackenburg and Margetts, 2024] Kobi Hackenburg and Helen Margetts. Evaluating the persuasive influence of political microtargeting with large language models. *Proceedings of the National Academy of Sciences*, 121(24):e2403116121, 2024.

[Hackenburg *et al.*, 2025] Kobi Hackenburg, Ben M Tappin, Luke Hewitt, Ed Saunders, Sid Black, Hause Lin, Catherine Fist, Helen Margetts, David G Rand, and Christopher Summerfield. The levers of political persuasion with conversational artificial intelligence. *Science*, 390(6777):eaea3884, 2025.

[He *et al.*, 2021] Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021.

[Huang *et al.*, 2025] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

[Ji *et al.*, 2024] Ziwei Ji, Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. Anah: Analytical annotation of hallucinations in large language models. *arXiv preprint arXiv:2405.20315*, 2024.

[Jiang *et al.*, 2021] Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021.

[Kalai *et al.*, 2026] Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. Evaluating large language models for accuracy incentivizes hallucinations. *Nature*, pages 1–3, 2026.

[Lewis *et al.*, 2020] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[Lin *et al.*, 2024] Zichao Lin, Shuyan Guan, Wending Zhang, Huiyan Zhang, Yugang Li, and Huaping Zhang. Towards trustworthy LLMs: A review on debiasing and dehallucinating in large language models. *Artificial Intelligence Review*, 57(9):243, August 2024.

[Lin *et al.*, 2025] Hause Lin, Gabriela Czarnek, Benjamin Lewis, Joshua P White, Adam J Berinsky, Thomas Costello, Gordon Pennycook, and David G Rand. Persuading voters using human–artificial intelligence dialogues. *Nature*, pages 1–8, 2025.

[Motoki *et al.*, 2025] Fabio Y. S. Motoki, Valdemar Pinho Neto, and Victor Rangel. Assessing political bias and value misalignment in generative artificial intelligence. *Journal of Economic Behavior & Organization*, page 106904, February 2025.

[Navigli *et al.*, 2023] Roberto Navigli, Simone Conia, and Björn Ross. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21, 2023.

[Niu *et al.*, 2024] Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models, May 2024.

[Noels *et al.*, 2025] Sander Noels, Guillaume Bied, Maarten Buyl, Alexander Rogiers, Yousra Fettach, Jefrey Lijffijt, and Tijl De Bie. What large language models do not talk about: An empirical study of moderation and censorship practices, 2025.

[Rettenberger *et al.*, 2025] Luca Rettenberger, Markus Reischl, and Mark Schutera. Assessing political bias in large language models. *Journal of Computational Social Science*, 8(2):42, May 2025.

[Rozado, 2025] David Rozado. Measuring Political Preferences in AI Systems: An Integrative Approach, March 2025.

[Santurkar *et al.*, 2023] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose Opinions Do Language Models Reflect?, March 2023.

[Vosoughi *et al.*, 2018] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018.

[Weidinger *et al.*, 2021] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*, 2021.

[Zhang *et al.*, 2023] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models, September 2023.